La cosa interessante è che ci sono sottoinsiemi con diversi livelli di agreement tra annotatori umani.

Può quindi fare uno studio replicando l'uso dei metodi di machine learning sui diversi sottoinsiemi per verificare l'ipotesi che la qualità dei dati di input determina la qualità delle predizioni.

IMPORTANTE!!!

20/30% da Sentences\_Merge.txt test set

il resto da Sentences\_66Agree.txt, Sentences\_75Agree.txt ( escludere le sentences presenti in AllAgree.txt) per creare vari training\_set

Ci siamo occupati di decidere quali dei vari dataset utilizzare per trainare il modello, validarlo e testarlo. Abbiamo in un primo momento verificato la distribuzione dei vari file, accertandoci che siano bilanciati allo stesso modo e abbiamo poi optato per il file All agree come training set (dato è il file più sicuro da utilizzare) mentre abbiamo usato le sentences rimanenti( all agree- 50 agree) per creare il nostro test set (30%) che dovrebbe essere composto da 1454 sentences.

Per il validation abbiamo invece optato per le sentences di 75 agree che non figurano nel file all agree: la scelta di questo dataset è stata fatta in quanto

Training Set:

61% neutral, 13% negative e 25% positive

Validation Set:

63 neutral/ 27 positive/ 10 negative

Test set

53% neutral, 34% positive, 13% negative

Test set: 1393

739 neutrali, 474 positive, 180 negativi

nuovo test set: 1267

739 neutrali, 348 positivo, 180 negativi

togli 126

Finetuning simpletransforming

rimozione di feature